Problems in Computational Biology:

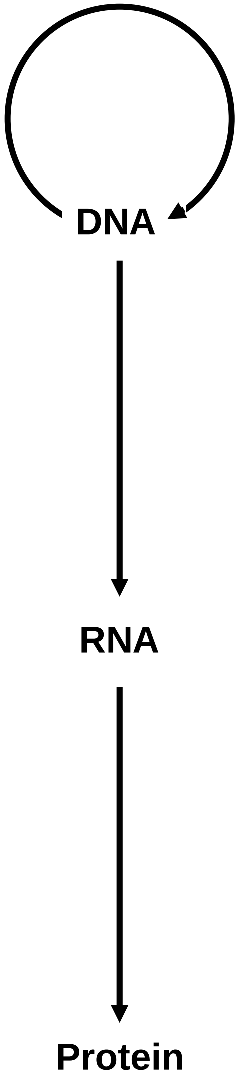
(1) Deciphering the Protein Complex Network

And maybe a little bit of:

(2) The Shape of RNA
or
(3) Finding non-coding RNA Genes

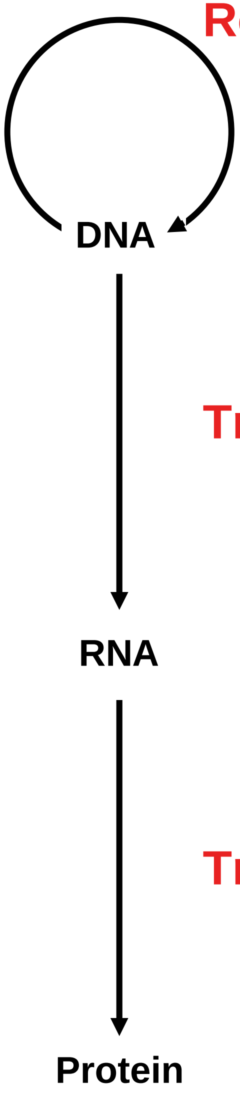
Richard F. Meraz
Lawrence Berkeley National Laboratory
Berkeley, CA

The Central Dogma



The Central Dogma

(gene) ATGAGTAACGCG
TACTCATTGCGC



Replication

duplication of DNA using DNA as the template

ATGAGTAACGCG
TACTCATTGCGC
+
(nontemplate, antisense) ATGAGTAACGCG
(template, sense) TACTCATTGCGC

Transcription

synthesis of RNA using DNA as the template

(mRNA) AUGAGUAACGCG
codon

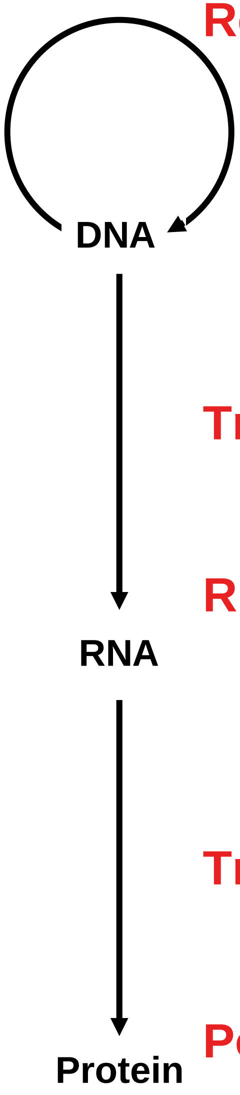
tRNA
ribosomes

Translation

(protein) MetSerAsnAla

synthesis of proteins using RNA as the template

The Central Dogma



Replication

1. DNA pol α and δ

Repair and recombination

2. DNA pol β and ε

Transcription

1. RNA pol I-ribosomal RNA (rRNA)
2. RNA pol II-messenger RNA (mRNA)
3. RNA pol III-5S rRNA, snRNA, tRNA

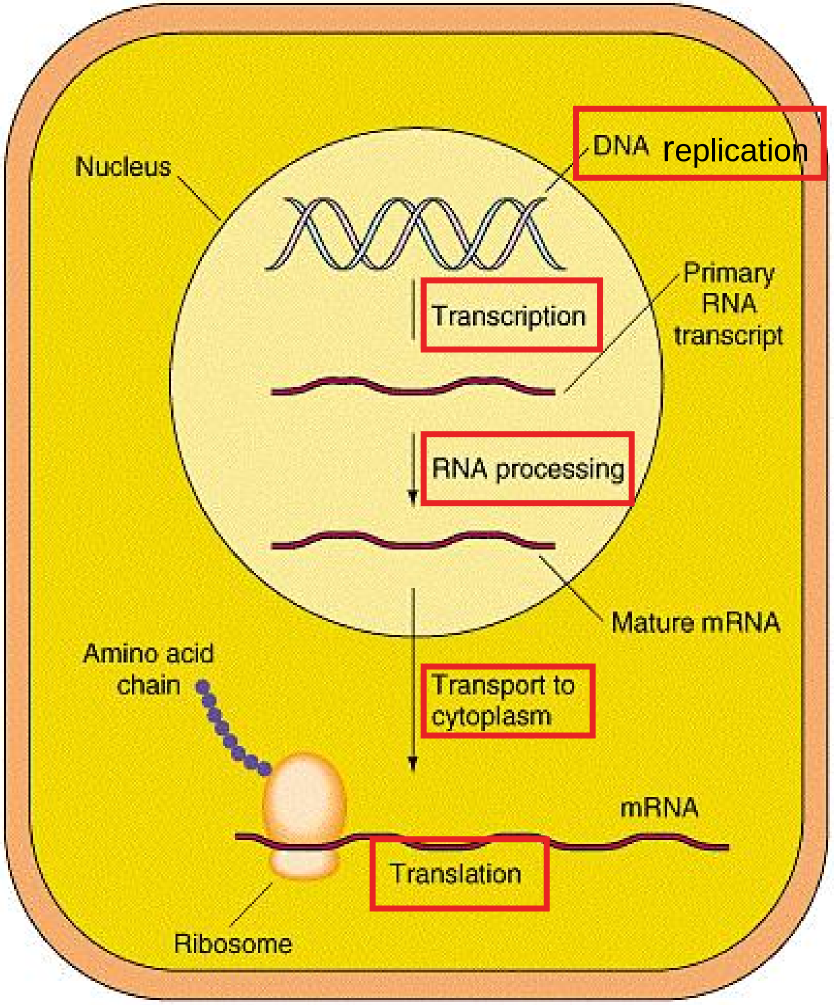
RNA processing

1. mRNA splicing
2. rRNA and tRNA processing
3. capping and polyadenylation

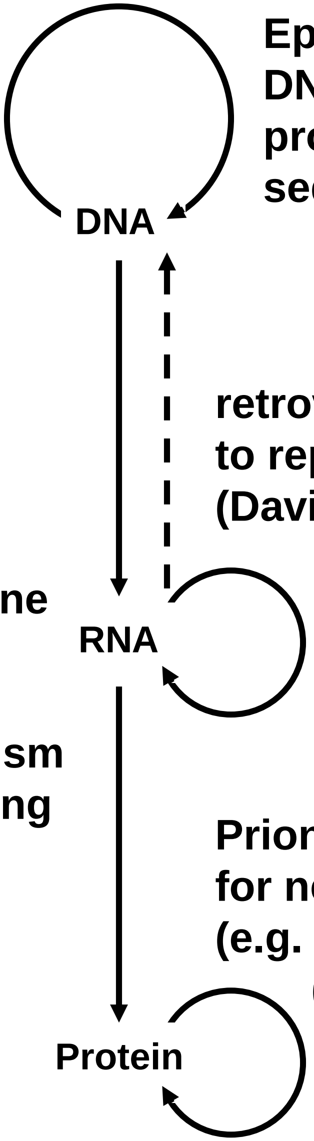
Translation

Post-translational modification

1. phosphorylatiotn
2. methylation
3. ubiquitination


Nucleus
DNA replication
Transcription
Primary RNA transcript
RNA processing
Mature mRNA
Amino acid chain
Transport to cytoplasm
mRNA
Translation
Ribosome

Exceptions to the Central Dogma (get Nobel Prizes)



Epigenetic marks, such as patterns of DNA methylation, can be inherited and provide information other than the DNA sequence.

retroviruses use reverse transcriptase to replicate their genome
(David Baltimore and Howard Temin)

RNA viruses

Prions are heritable proteins responsible for neurological infectious diseases
(e.g. scrapie and mad cow)
(Stanley Pruisner)

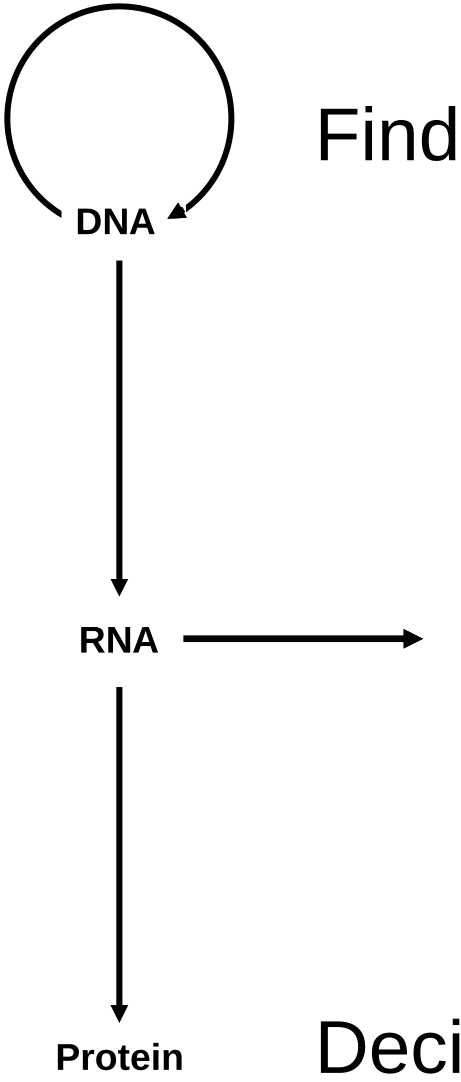
mRNA introns (splicing)
(Philip Sharp and Richard Roberts)

RNA editing (deamination of cytosine to yield uracil in mRNA)

RNA interference (RNAi) a mechanism of post-transcriptional gene silencing utilizing double-stranded RNA

RNAs (ribozymes) can catalyze an enzymatic reaction
(Thomas Cech and Sidney Altman)

The Central Dogma



Finding non-coding RNA Genes

Non-coding or functional RNA

The Shape of (non-coding) RNA

Deciphering the Protein Complex Network

Genome: 30.000 genes

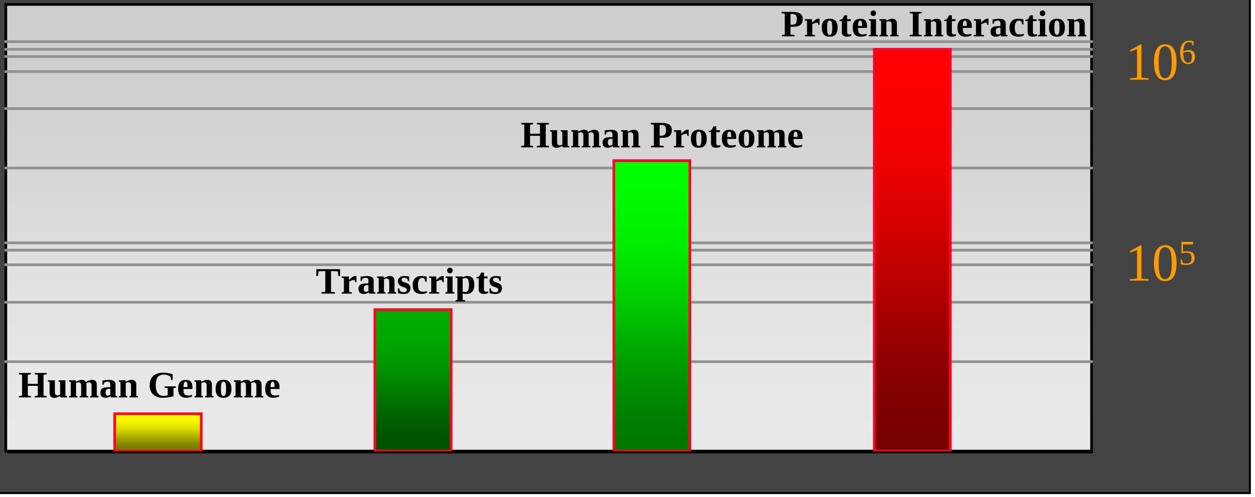
↓

Transcriptome: 40-100.000 mRNAs

↓

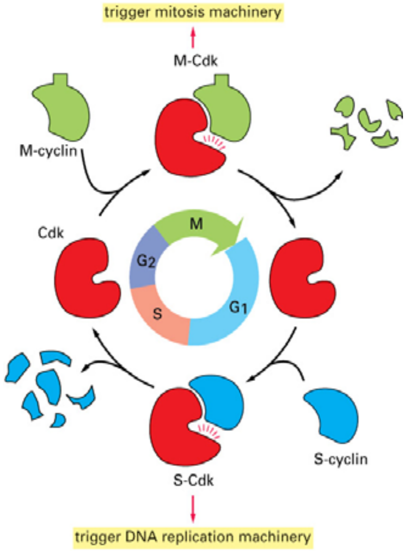
Proteome: 100-400.000 proteins

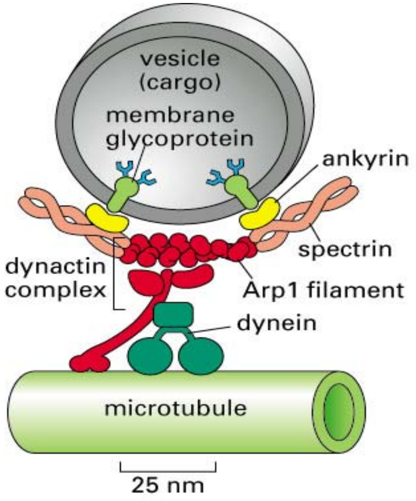
>1.000.000 interactions



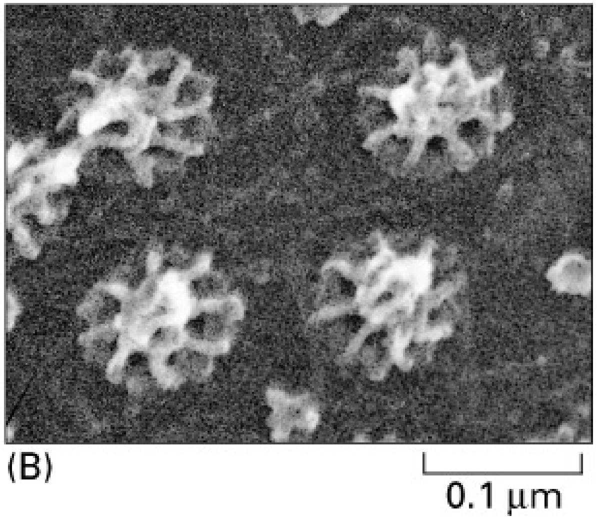
Protein-Complexes – Who Cares ?

- **Protein complexes are important for virtually every biological process and most diseases.**





- **Genome sequences identify tens of thousands of genes: linking these to 200-300 core biological processes will make their study manageable.**



High-throughput methods for detecting Protein Complexes

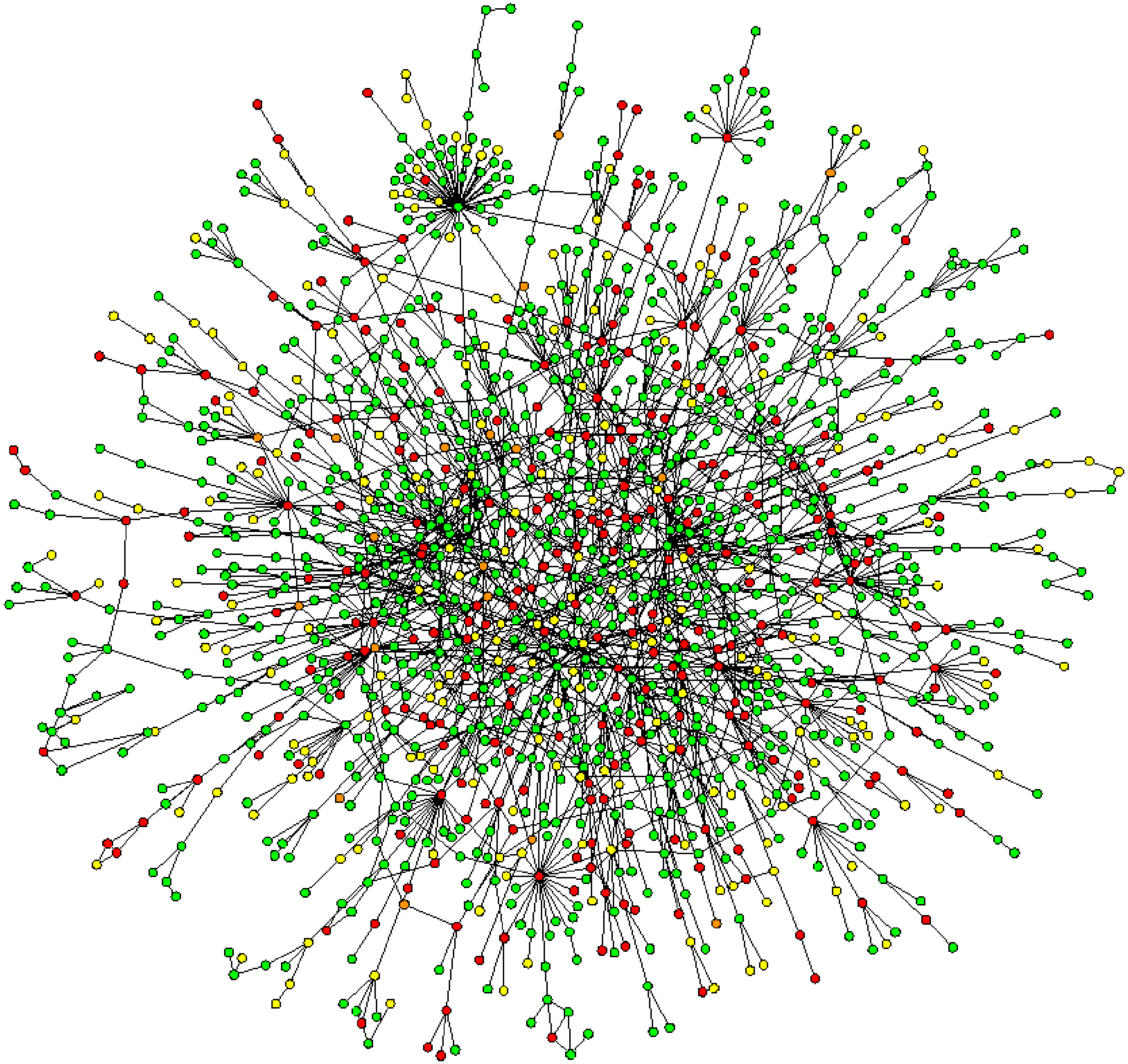
- Two-hybrid dataset by Uetz *et al* 2000 (the first comprehensive study in yeast)
- Two-hybrid dataset by Ito *et al* 2001 (broad coverage in yeast)
- HMS-PCI dataset by Ho *et al* 2002
- TAP-MS dataset by Gavin *et al* 2002

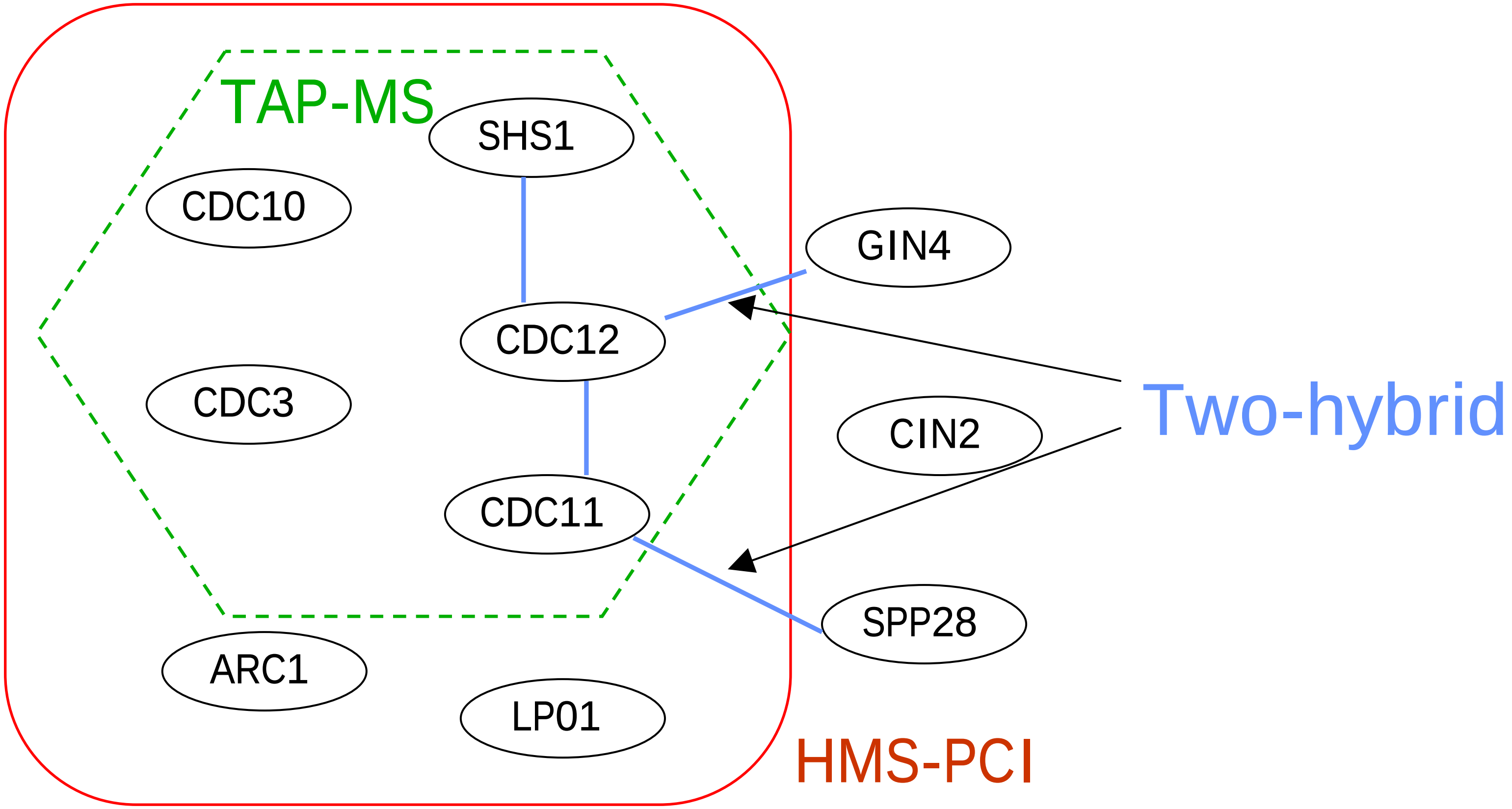
Very little overlap between interaction data from different experiments.

	ITO *et al*	**Uetz *et al***	**Gavin *et al***	**Ho *et al***
Ito *et al*	4363	**186**	**54**	**63**
Uetz *et al*	**186**	1403	**54**	**56**
Gavin *et al*	**54**	**54**	3222	**198**
Ho *et al*	**63**	**56**	**198**	3596

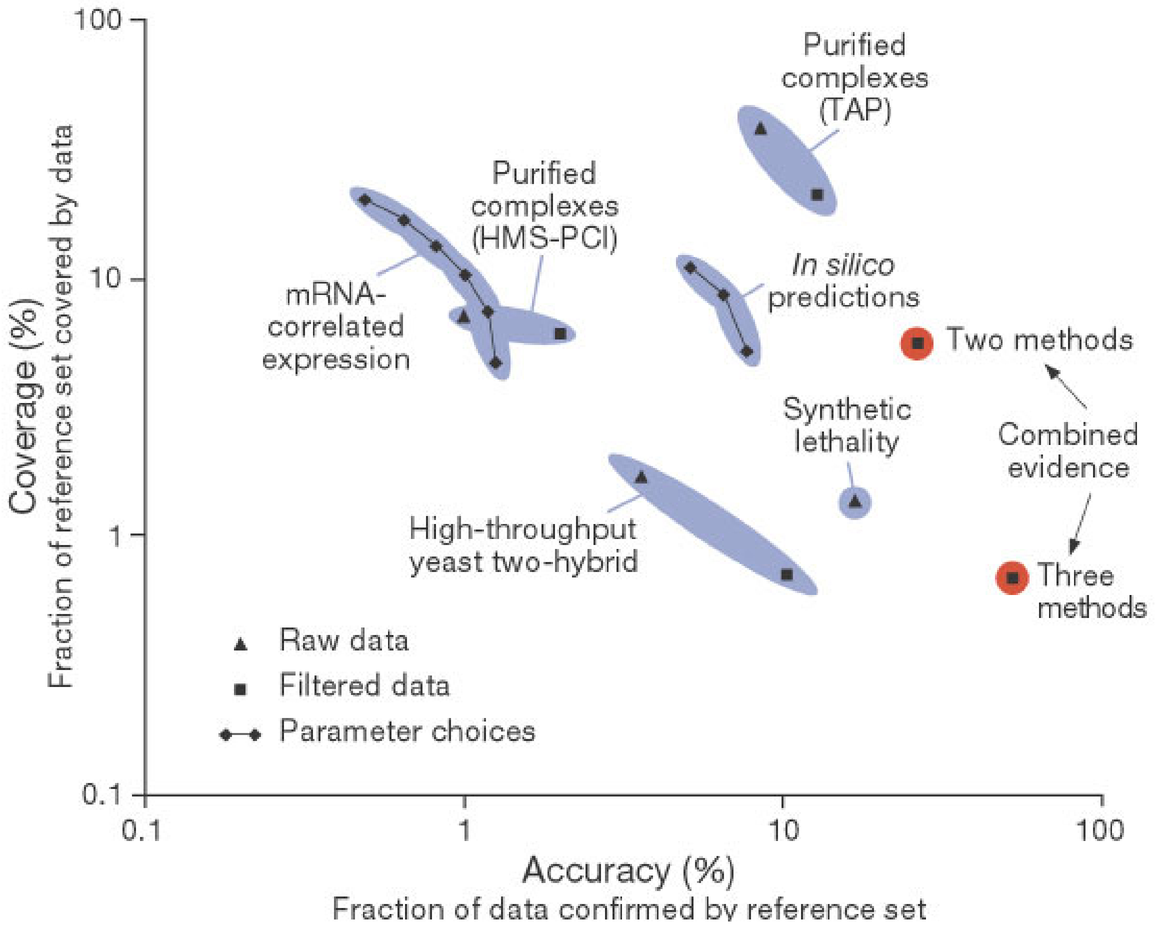
Copied from Salwinski and Eisenberg, 2003

Interactions in the yeast proteome

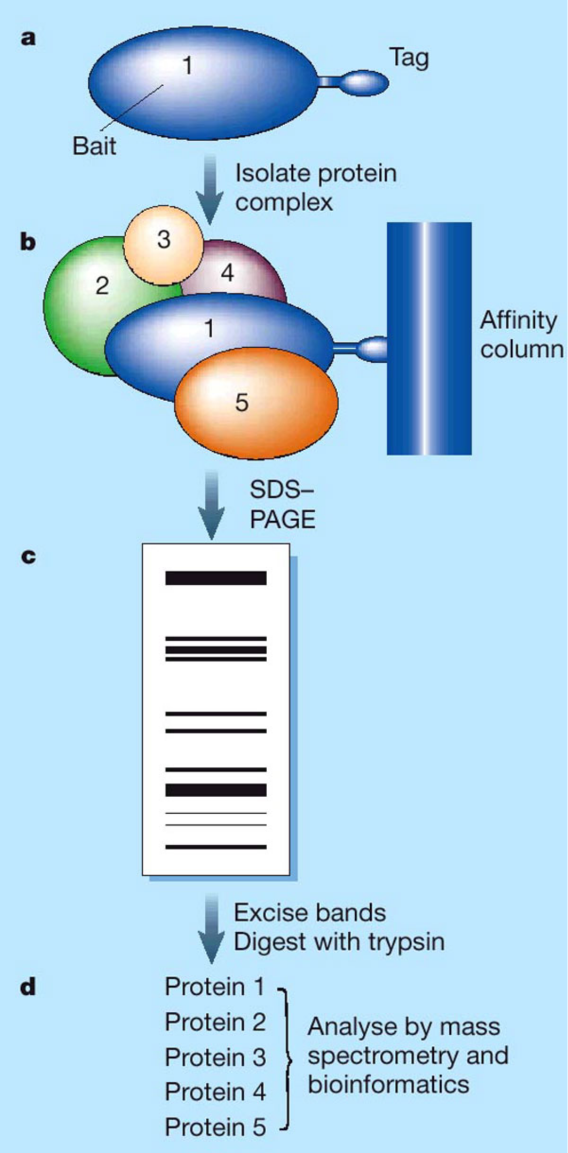



TAP-MS
SHS1
CDC10
GIN4
CDC12
CDC3
Two-hybrid
CIN2
CDC11
SPP28
ARC1
LP01
HMS-PCI

Not all data-sets are created equal.



von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. Comparative assessment of large-scale data sets of protein-protein interactions. Nature 2002;417(6887):399-403.

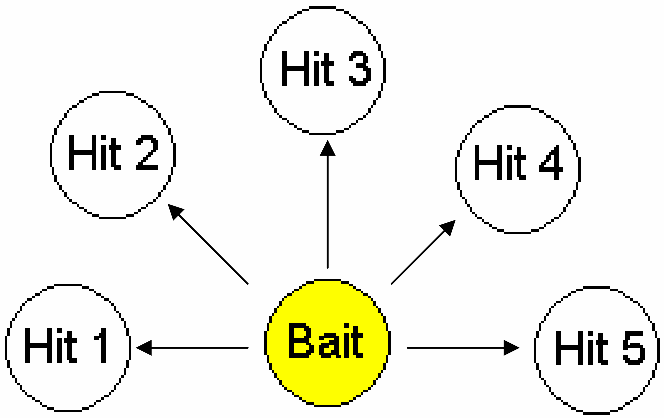


- Tandem-Affinity Purification coupled with Mass-Spectrometry (TAP-MS) determines the constituents of multi-protein complexes.

Gavin AC, *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002;415(6868):141-147.

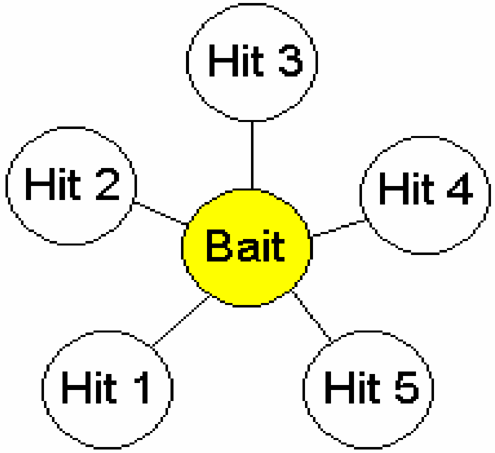
Computer Scientists who don't understand biology make bad assumptions

Data

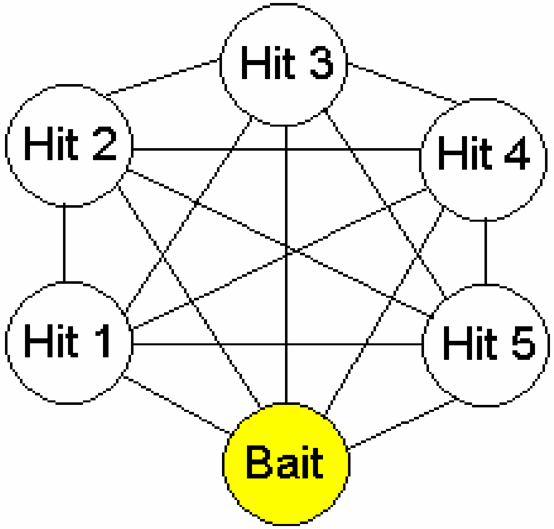


K-cores (Bader and Hogue, 2002)
Cliques (Spirin and Mirny, 2003)
Hypergraph – *k*-core (Pothen, 2003)

Spoke Model



Matrix Model



Looking at the Network

- Reduce noise by eliminating unnecessary assumptions about which proteins interact

- Do not ignore that fact that proteins that coincide in more than one complex are likely to somehow integrate their functions.

- Don’t ignore the notion of ‘communication’ and coupling that occurs when protein complexes share components. This is the higher-level organization of the network via linking of biological processes.

Goals + Hypothesis

- Unify two notions of network – proteins interacting, but also complexes interacting via the notion of ‘shared components’.
- Partition these networks into ‘modules’ or functional units separable from the rest of the network. This is the goal of systems-level or network biology.
- The framework should aide in reasoning about uncharacterized protein components and be biologically consistent.

A Unified Representation of Multi-Protein Complex Networks

Dual relationship between protein and protein-complex is specified by adjacency matrix ***B***.

Protein-Protein (p-p) interaction network:

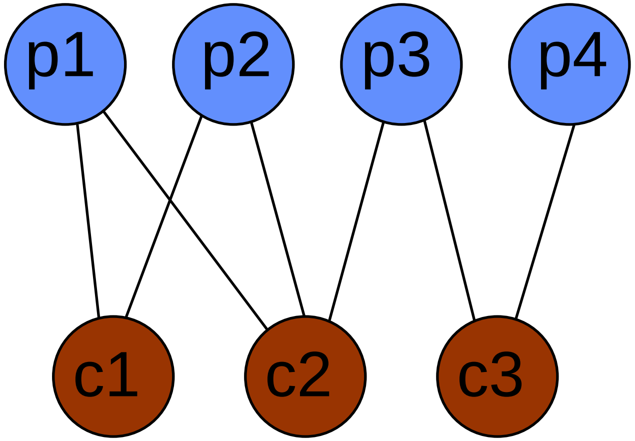
$$(\mathbf{BB}^T)_{ij} = \left(\begin{array}{c} \#\text{ of protein complexes} \\ \text{containing both proteins } p_i, p_j \end{array} \right)$$

Complex-Complex (c-c) interaction network

$$(\mathbf{B}^{\mathrm{T}}\mathbf{B})_{ij} = \left(\begin{array}{c} \#\text{ of proteins shared by} \\ \text{protein complexes } c_i, c_j \end{array} \right)$$

Toy Protein Complex Dataset

Bipartite Graph



Adjacency Matrix

$\mathbf{B}$

	c1	c2	C3
P1	1	1	0
P2	1	1	0
P3	0	1	1
P4	0	0	1

$\mathbf{B}^T$

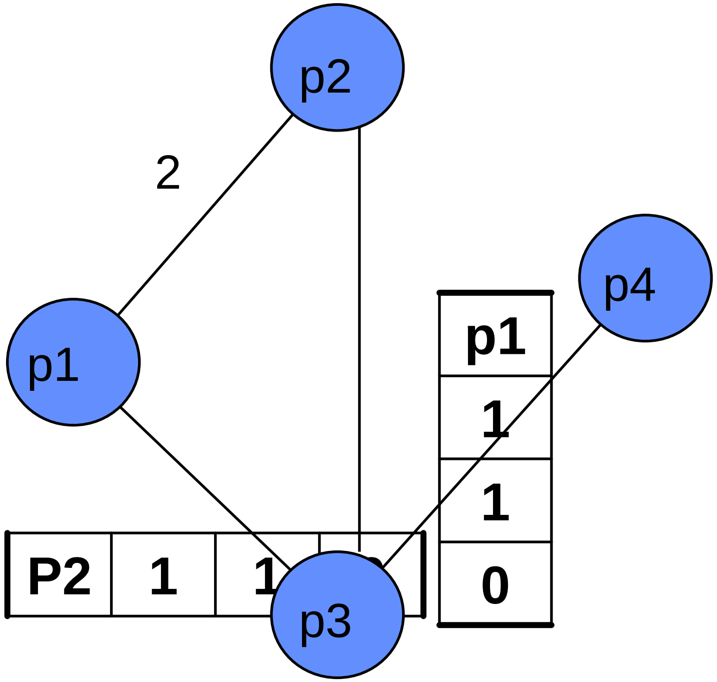
	p1	p2	p3	P4
C1	1	1	0	0
C2	1	1	1	0
C3	0	0	1	1

B

	c1	c2	C3
P1	1	1	0
P2	1	1	0
P3	0	1	1
P4	0	0	1

B^T

	p1	p2	p3	P4
C1	1	1	0	0
C2	1	1	1	0
C3	0	0	1	1



P2	1	1	[illegible]

p1
1
1
0

	p1	p2	p3	P4
P1	2	2	1	0
P2	2	2	1	0
P3	1	1	2	1
P4	0	0	1	1

Clustering P-P and C-C Network

$$s(G_1,G_2)=\sum_{i\in G_1}\sum_{j\in G_2} w_{ij}$$

Connectivity

$$J(G_1,G_2)=\frac{s(G_1,G_2)}{s(G_1,G_1)}+\frac{s(G_1,G_2)}{s(G_2,G_2)}$$

Cohesion between two graphs

$$q(i)=\begin{cases} a & \text{if } i\in G_1 \\ -b & \text{if } i\in G_2 \end{cases}$$

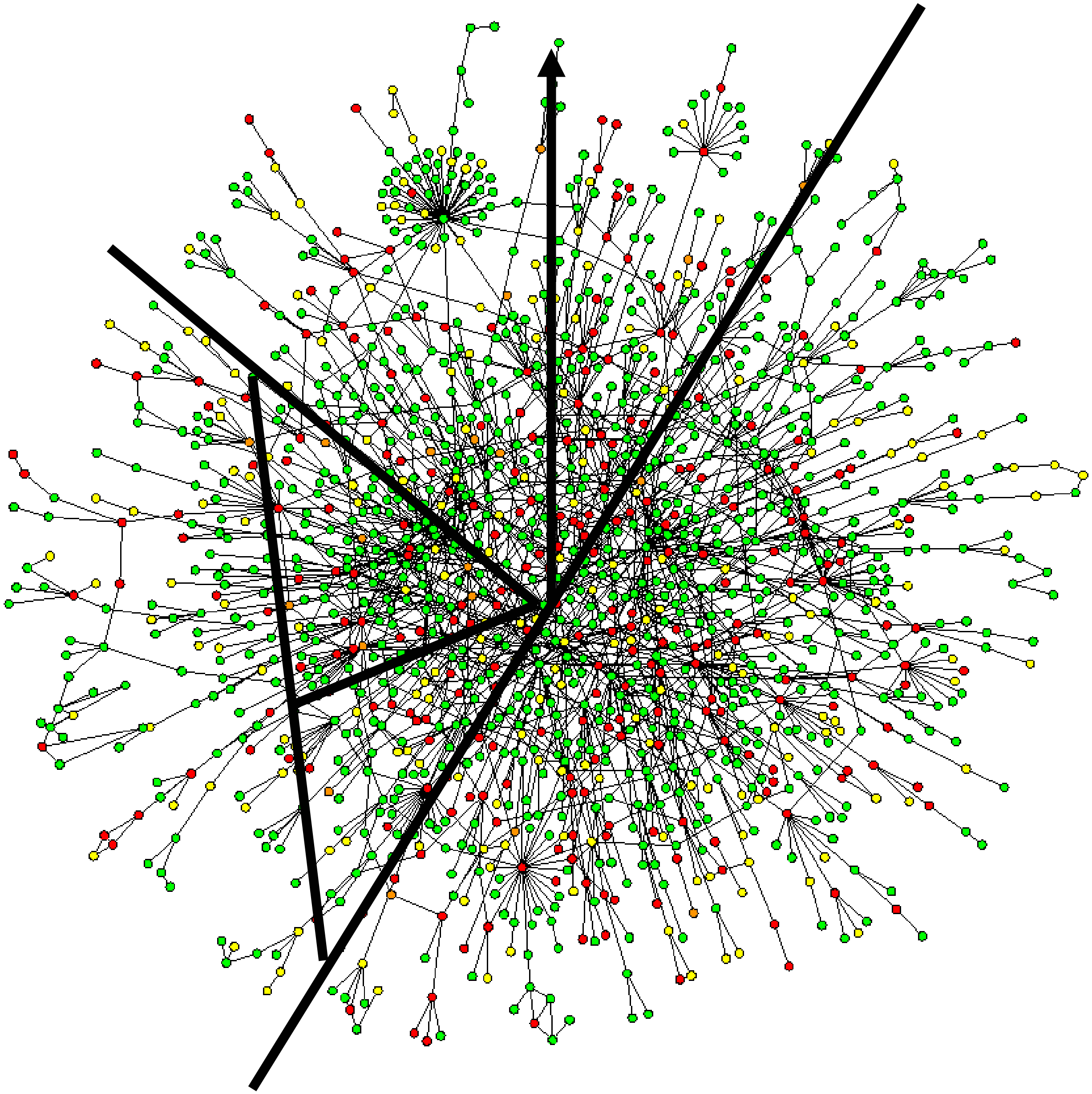
Write the solution like this

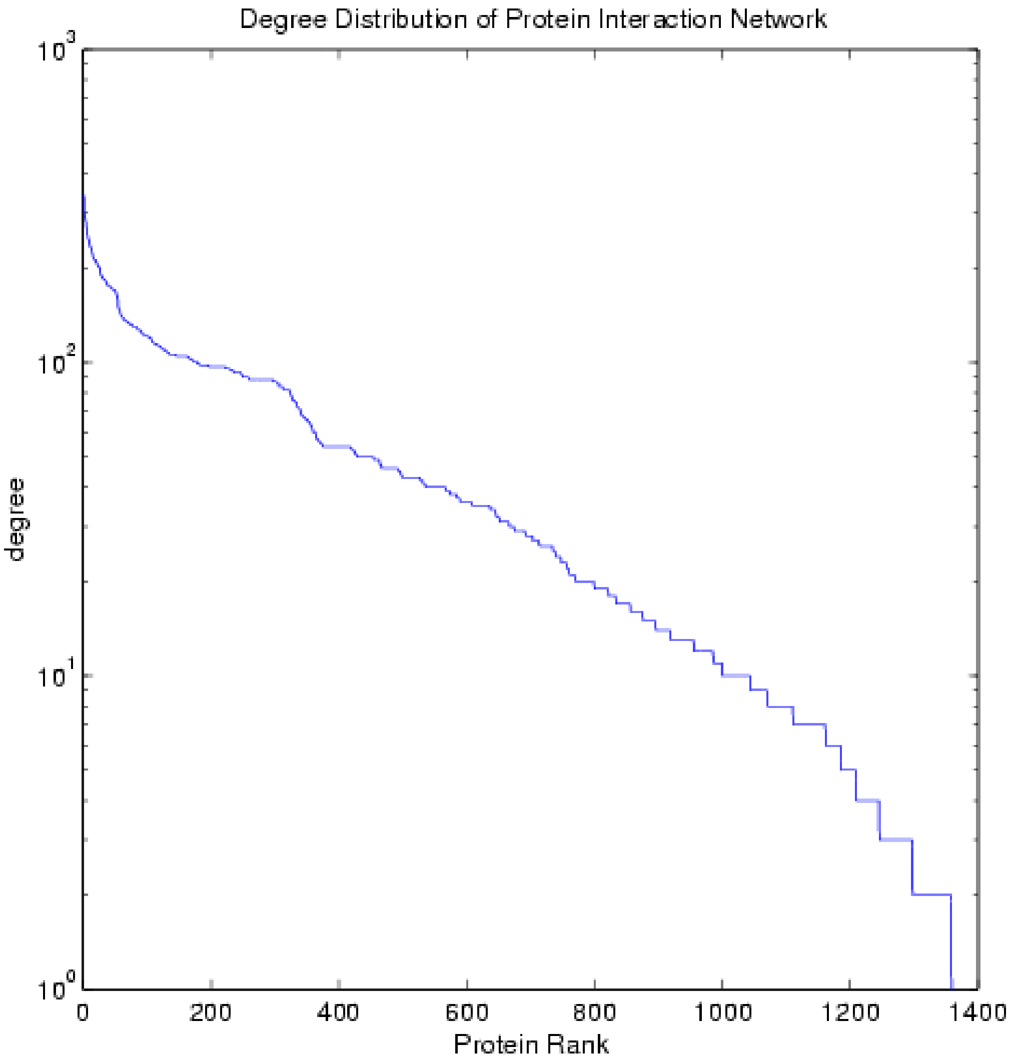
It follows (ie. Proof omitted) :

$$\min_{\mathbf{q}} J(G_1,G_2) \Rightarrow \min_{\mathbf{q}} \frac{\mathbf{q}^T(\mathbf{D}-\mathbf{W})\mathbf{q}}{\mathbf{q}^{\mathrm{T}}\mathbf{D}\mathbf{q}}$$

Solution is eigenvector corresponding to second smallest eigenvalue

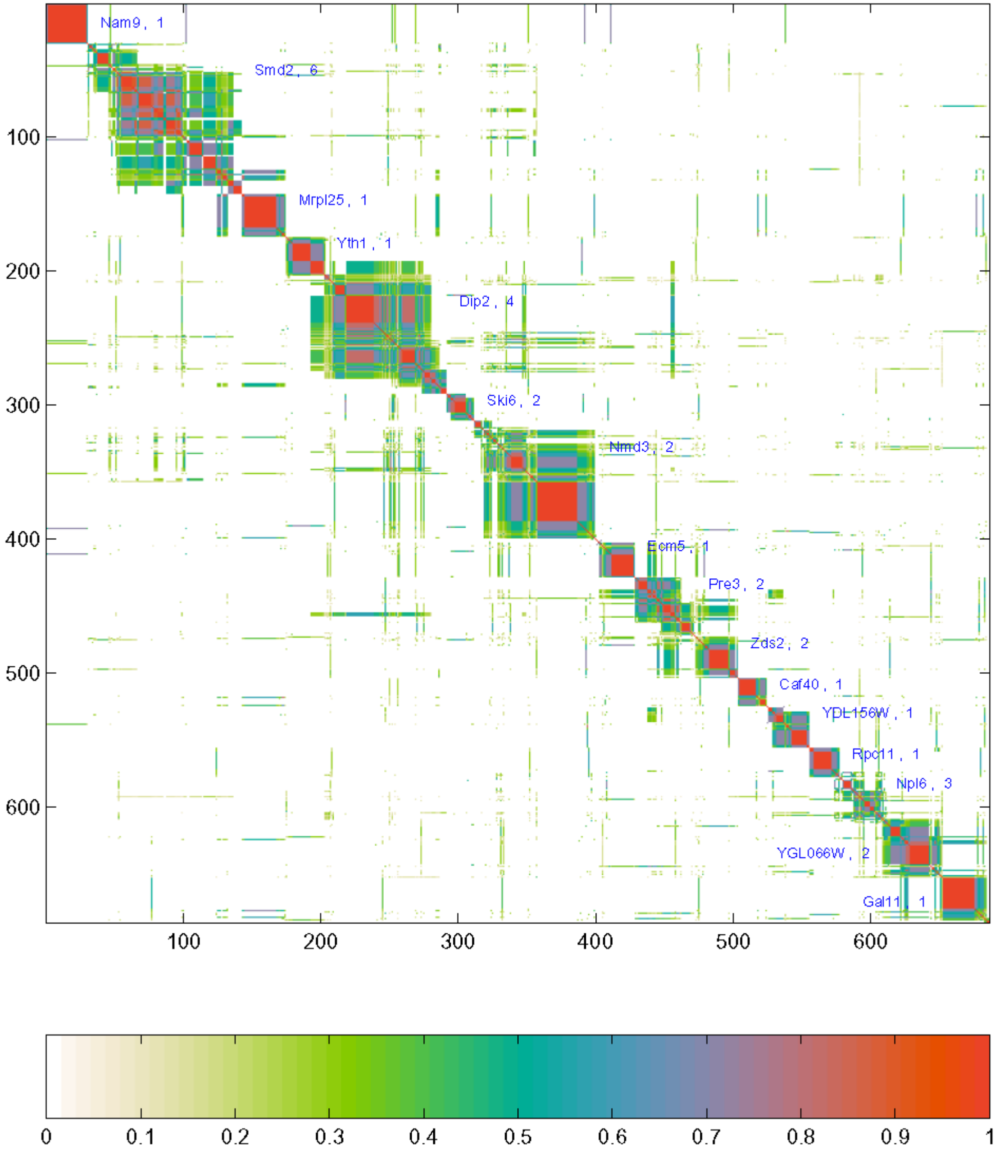
$$(\mathbf{D}-\mathbf{W})\mathbf{q}=\lambda\mathbf{D}\mathbf{q}$$



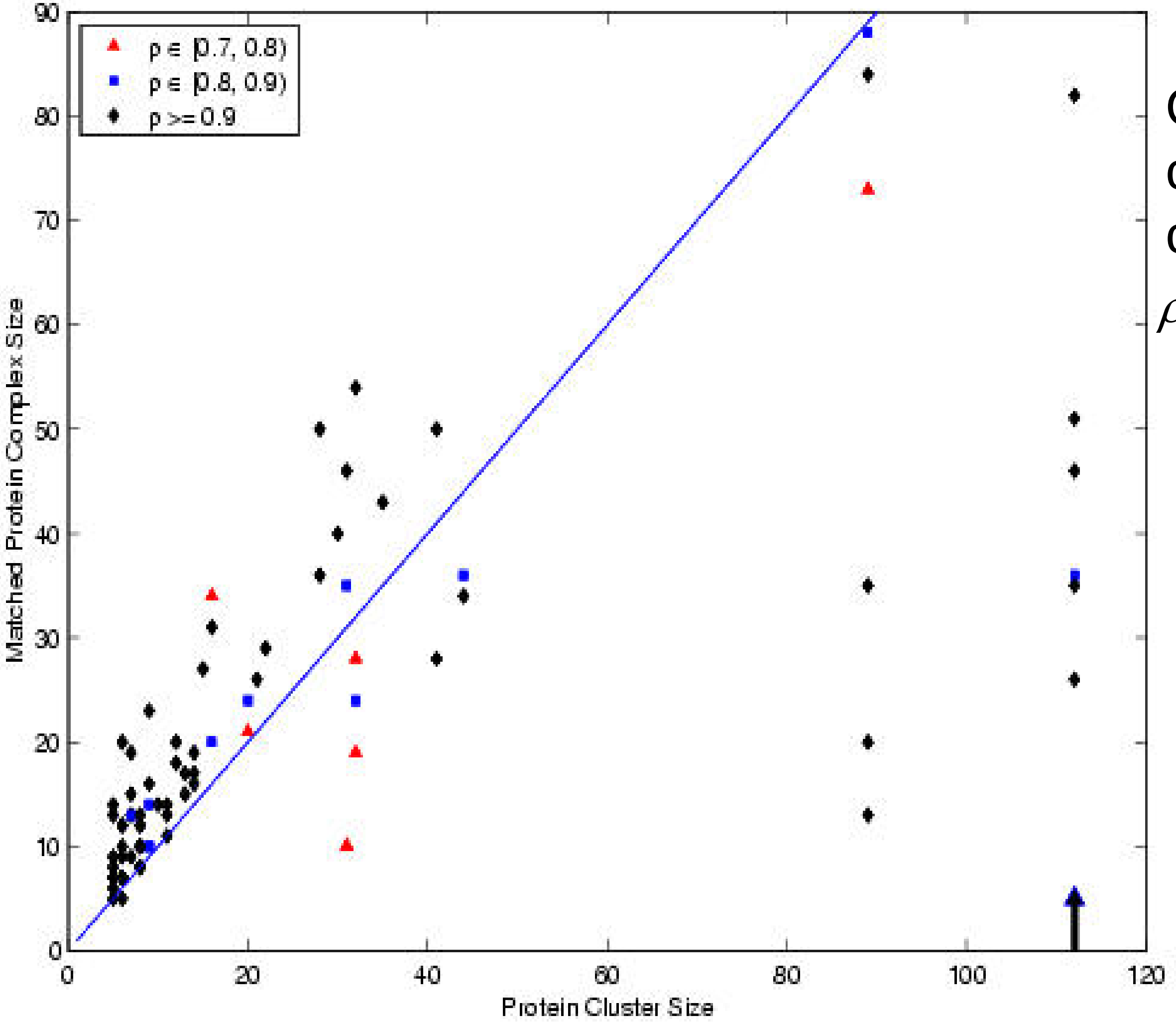

Degree Distribution of Protein Interaction Network
10^3
10^2
10^1
10^0
degree
0
200
400
600
800
1000
1200
1400
Protein Rank

Predicted Modules(clusters) of P-P network

Connections in the network are weighted by the number of complexes in which two proteins are coincident.



PP-modules overlap experimental complexes



Overlap between protein clusters and protein complexes defined as

$$\rho = n(P_k, c_j) / \min(|P_k|, |c_j|)$$

- Discovered protein clusters highly overlap with experimental complexes
- Uncharacterized proteins in discovered clusters might infer novel functions.

Modules in P-P network reflect characteristic physical and chemical properties

Lys	100	Asn	56	Val	30	Ile	24
Asp	89	Gln	50	Tyr	29	Ser	23
Arg	73	Cys	39	Met	29	Leu	22
Pro	70	His	33	Trp	28	Gly	21
Glu	66	Ala	31	Thr	28	Phe	21
pI	169	Basic	149	Acidic	97	MW	60
Aromatic	30	Helix	37	Beta-Sheet	33	Coil	27

$$F = \frac{1}{K-1}\sum_{k=1}^{K} n_k(\bar{f}_k - \bar{f}) / \frac{1}{n-K}\sum_{k=1}^{K}(n_k - 1)\sigma_k^2$$

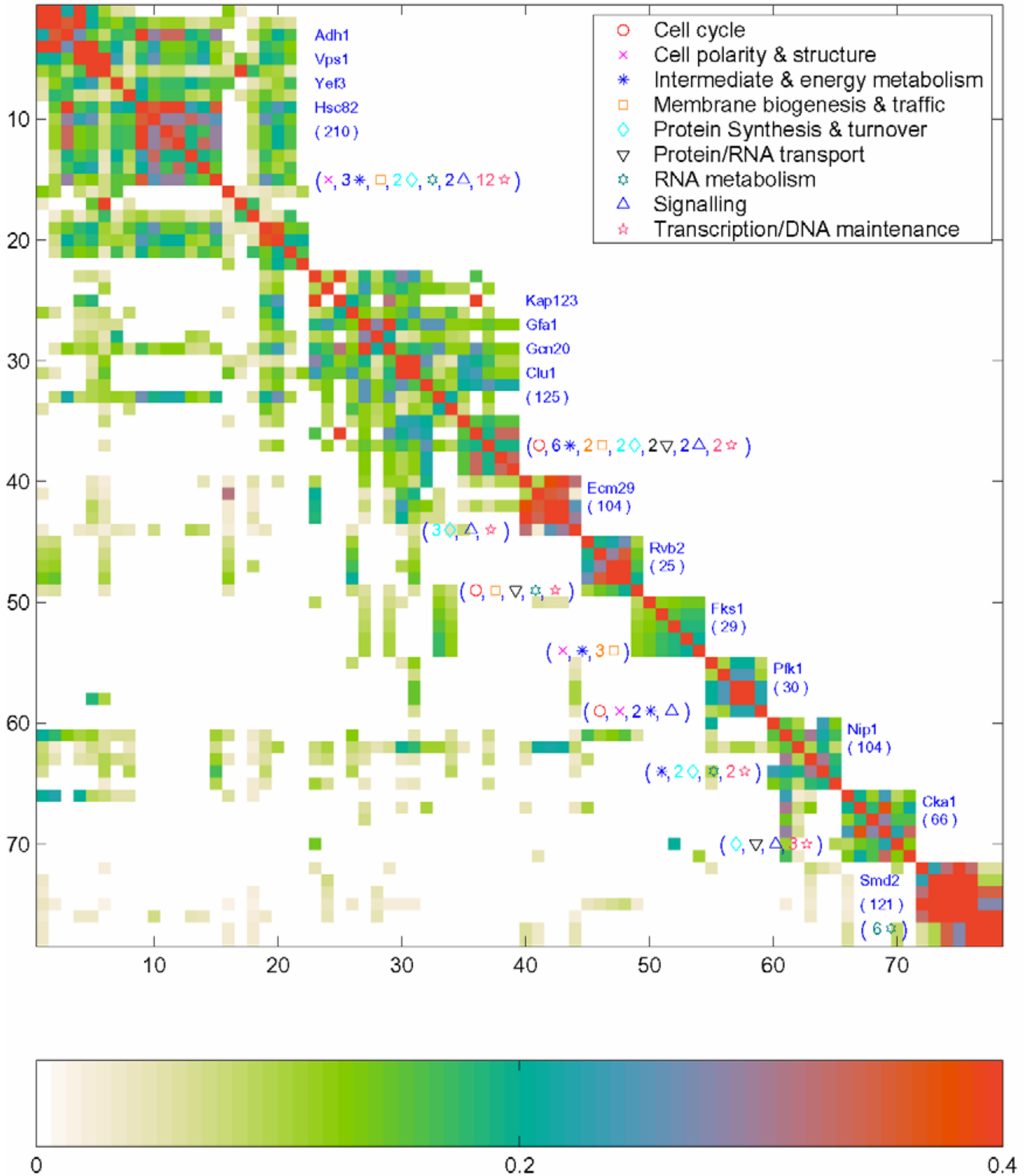
Polar residues (Lys, Arg, Gln, Asn,Asp), hydrogen bonding (Arg), hydrophobic interactions (Pro).

Covalent Modification (methylation and acetylation) of Arg and Lys

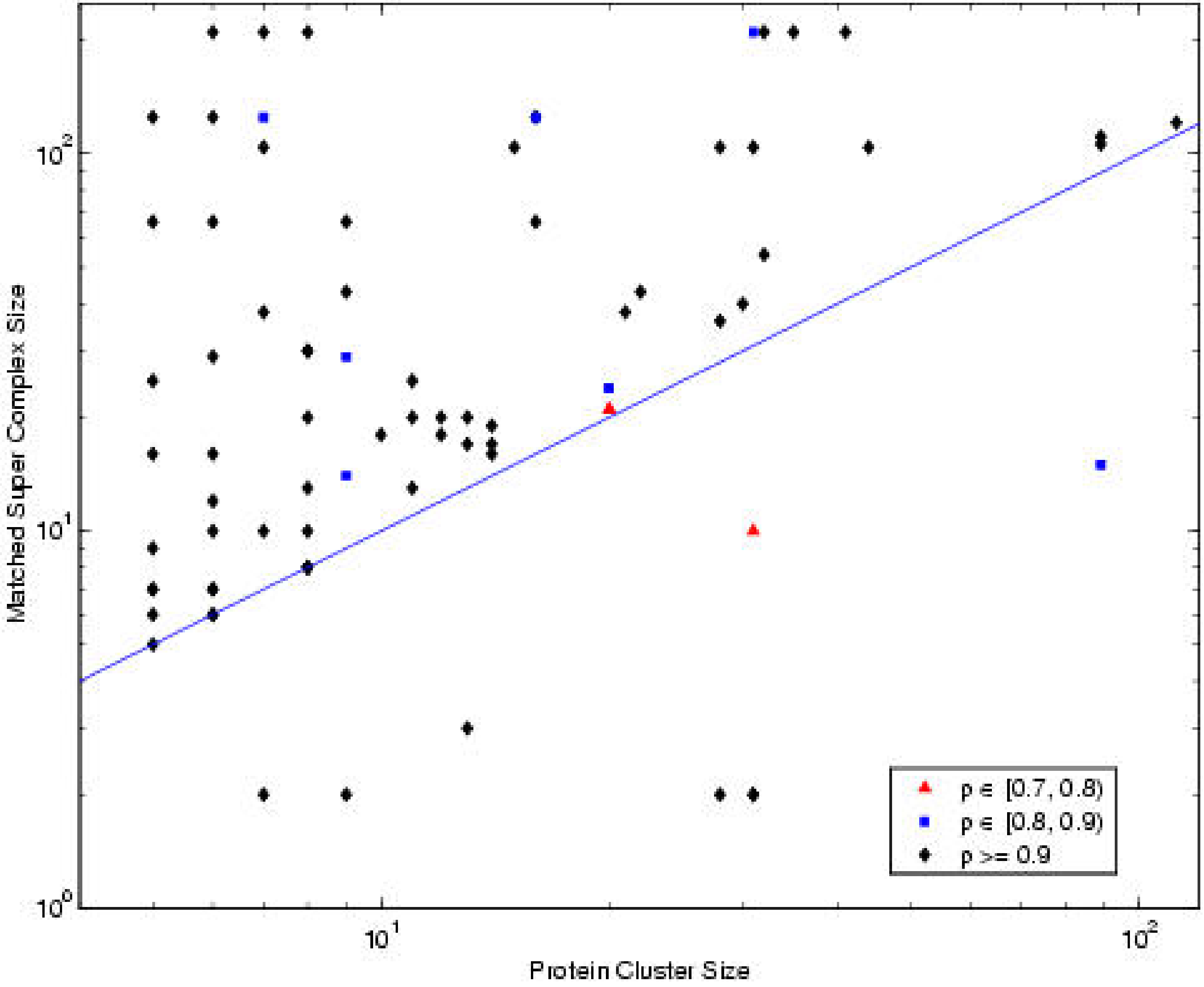
Disulfide bonds and cys.

Secondary structure features uniformly distributed at protein interaction interfaces.

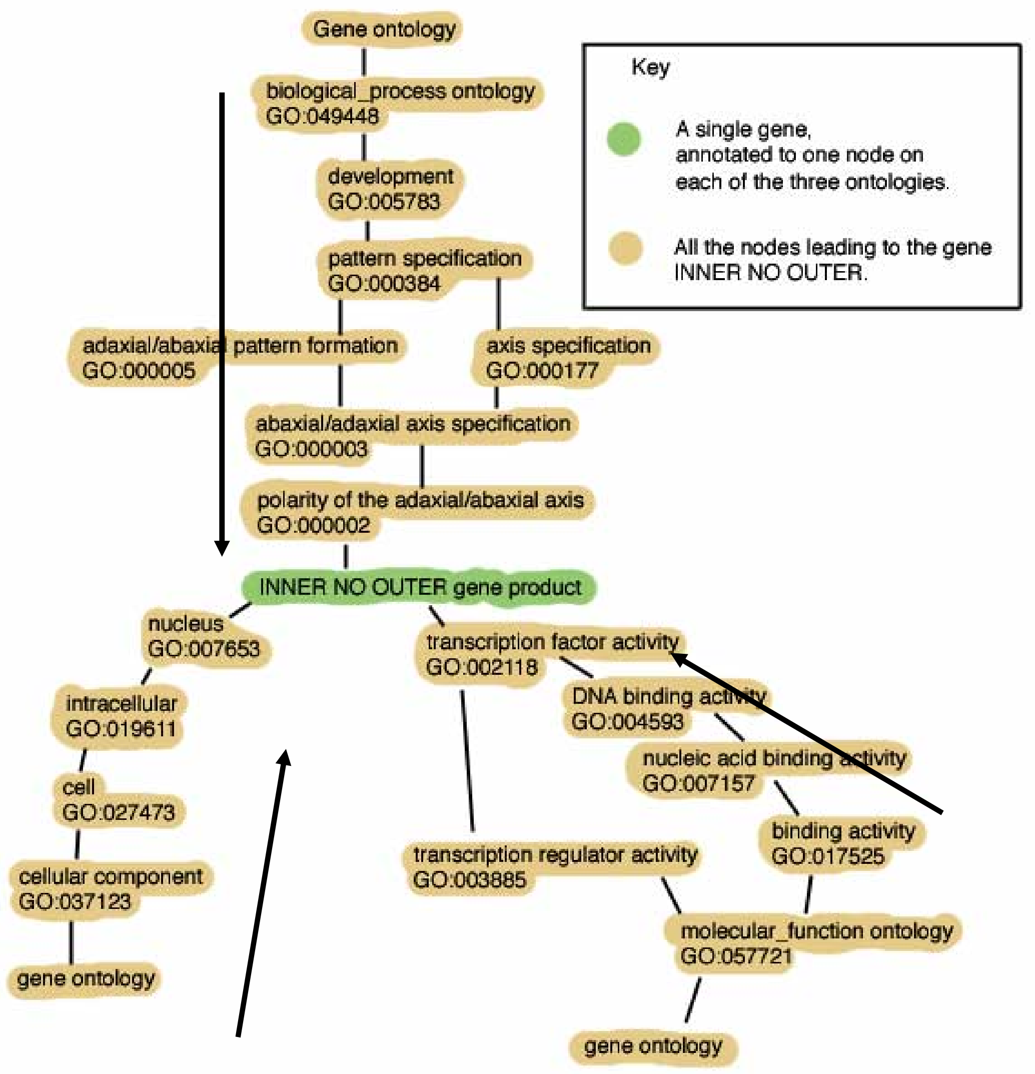
Connections in the network are weighted by the number of proteins that two complexes share.



Supercomplexes overlap modules in PP-network



Gene Ontology (GO)


Gene ontology
biological_process ontology GO:049448
development GO:005783
pattern specification GO:000384
adaxial/abaxial pattern formation GO:000005
axis specification GO:000177
abaxial/adaxial axis specification GO:000003
polarity of the adaxial/abaxial axis GO:000002
INNER NO OUTER gene product
nucleus GO:007653
intracellular GO:019611
cell GO:027473
cellular component GO:037123
gene ontology
transcription factor activity GO:002118
DNA binding activity GO:004593
nucleic acid binding activity GO:007157
binding activity GO:017525
transcription regulator activity GO:003885
molecular_function ontology GO:057721
gene ontology
Key
A single gene, annotated to one node on each of the three ontologies.
All the nodes leading to the gene INNER NO OUTER.

Three separate ontologies: Biological Process, Molecular Function, Cellular Component.

Organized as a DAG describing gene products (proteins and functional RNA).

Makes the represented biological relationships computable.

Collaborative effort between major genome databases.

http://www.geneontology.org

Gene Ontology

- Molecular function **catalytic or binding activities at (e.g. nucleic acid binding or exonuclease)**

- Biological process **is accomplished by ordered assemblies, pathways, with concerted function (e.g. ‘signal transduction’ or ‘nuclear export’).**

- Cellular component **compartmental, obligatory, or logical grouping (e.g. nucleus or spliceosome).**

Validating the Modules in the Networks

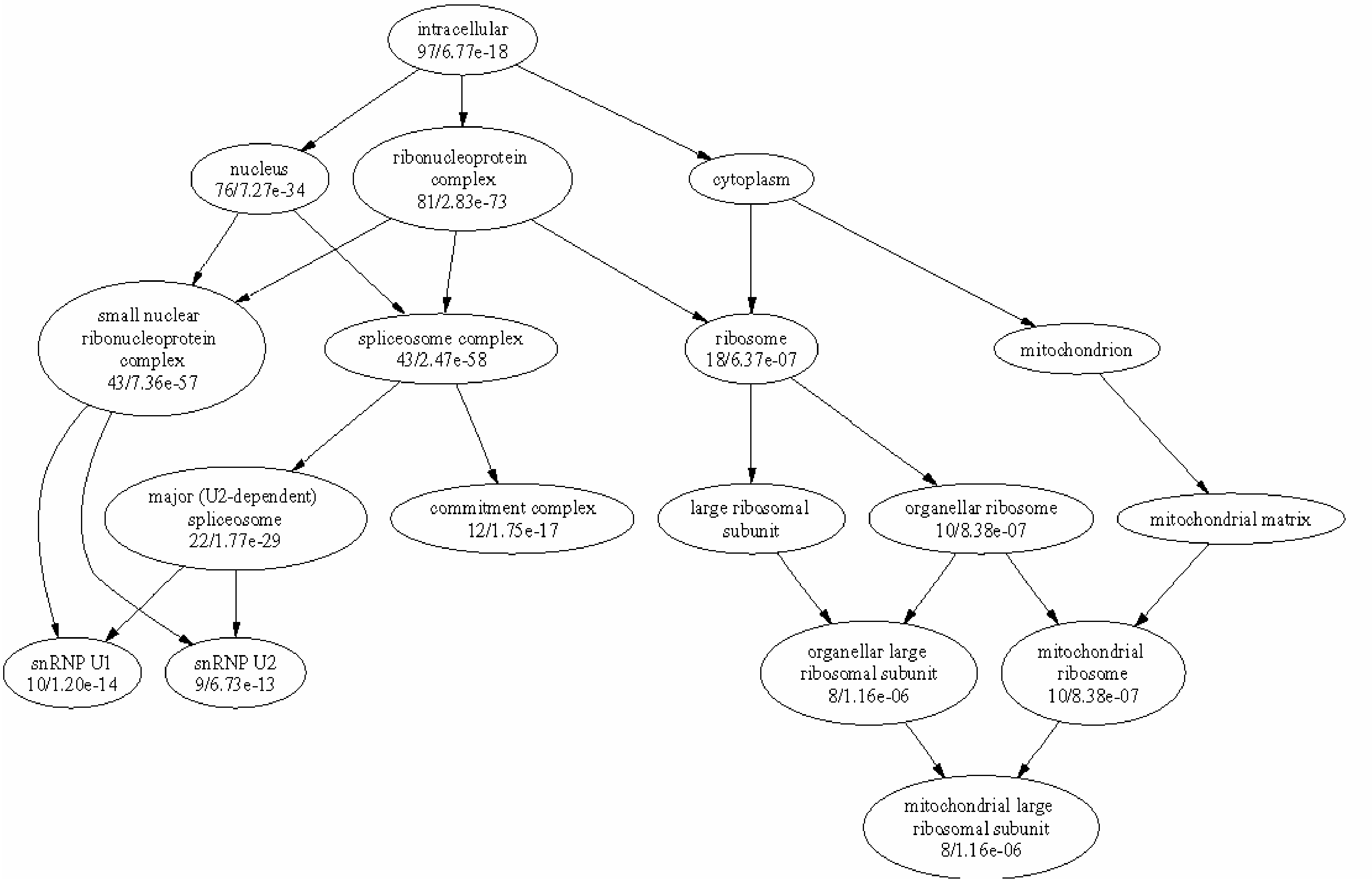
DAG structure of GO formalizes knowledge in biology by making it computable.

For constituent proteins in each cluster annotate To most specific term. Ascend the graph and annotate with parent terms.

Annotations observed by chance?

$$P = \sum_{n \le j \le N} \binom{N}{j} p^{j} (1 - p)^{N - j}$$

Computationally Discovered Modules are Biologically Consistent

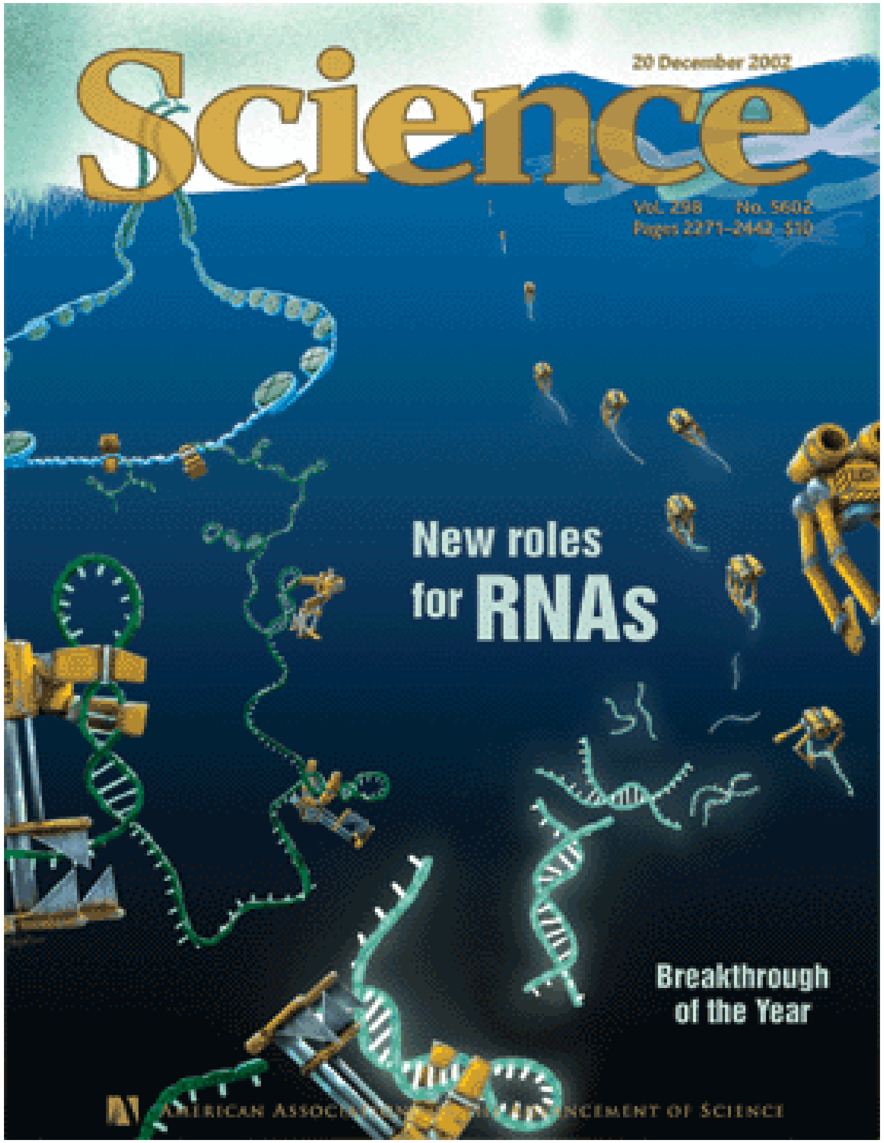

intracellular 97/6.77e-18
nucleus 76/7.27e-34
ribonucleoprotein complex 81/2.83e-73
cytoplasm
small nuclear ribonucleoprotein complex 43/7.36e-57
spliceosome complex 43/2.47e-58
ribosome 18/6.37e-07
mitochondrion
major (U2-dependent) spliceosome 22/1.77e-29
commitment complex 12/1.75e-17
large ribosomal subunit
organellar ribosome 10/8.38e-07
mitochondrial matrix
snRNP U1 10/1.20e-14
snRNP U2 9/6.73e-13
organellar large ribosomal subunit 8/1.16e-06
mitochondrial ribosome 10/8.38e-07
mitochondrial large ribosomal subunit 8/1.16e-06

Verified Complexes in Supercomplex 47

MIPS Annotation Category	# ORFs in C_{47}	# ORFs matched
RNA Pol II holoenzyme	35	23
Kornberg's mediator	21	21
Other transcription	73	17
HAT A	15	14
TFIID	13	13
SAGA	14	13
Ada-Spt	14	13
TAFIIs	12	12
DNA repair	33	9
RSC	10	6
ADA	6	6
Replication fork	30	6
DNA mismatch repair	5	5
Cytoplasmic translation initiation	27	4
SAGA-like	5	4
Nucleotide excision repairosome	16	3
RNA Polymerase III	13	3
Replication factor A	3	3
Actin-associated motorproteins	7	3
MSH2/MSH3	3	3
Srb10p	4	3
NEF4	2	2
eIF4A	2	2
NuA4	2	2
Nuclear pore	24	2
Sir	2	2

- Transcription
- Gene silencing
- Replication
- RNA processing
- RNA modification
- RNA stability
- mRNA translation
- Protein stability
- Protein translocation
- Metabolite sensing and regulation

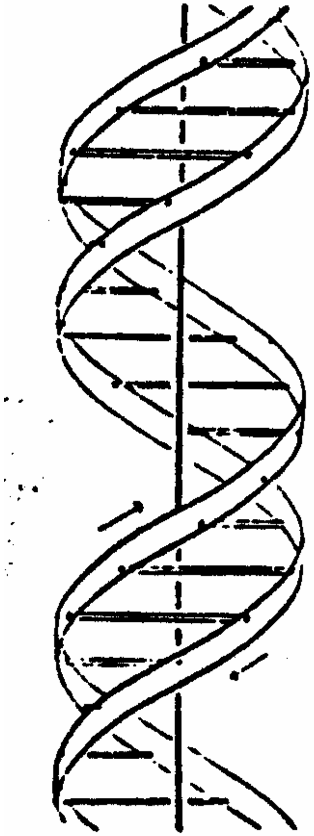
The Number of Known Functional RNAs in E. coli has Grown from 10 to 70 since 2000.



The genome, famously, is digital

1892: Miescher postulates that genetic information may be encoded in a linear form using a few different chemical units:

“…just as all the words and concepts in all languages can find expression in twenty-four to thirty letters of the alphabet.”



This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

Symbolic texts can be cracked

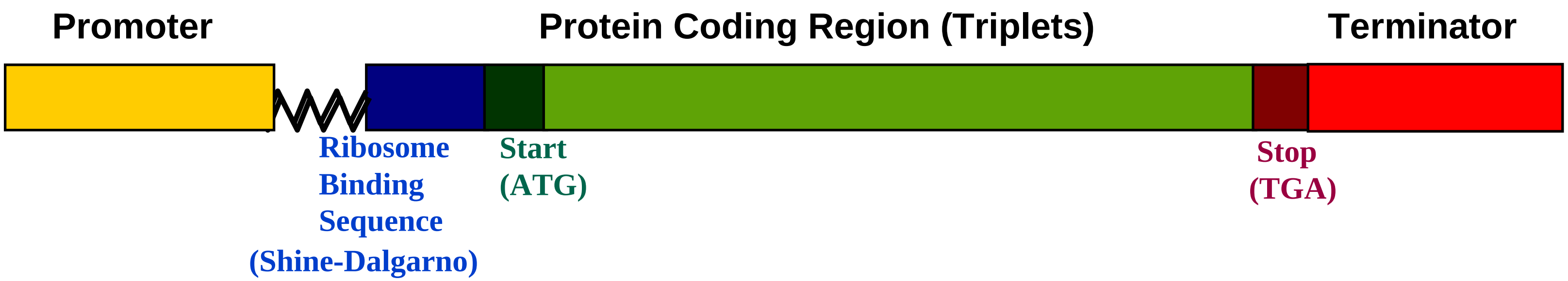
Michael Ventris and John Chadwick, 1953

"Cryptography has contributed a new weapon to the student of unknown scripts.... the basic principle is the analysis and indexing of coded texts, so that underlying patterns and regularities can be discovered. ***If a number of instances can be collected***, ***it may appear that a certain group of signs in the coded text has a particular function***...."

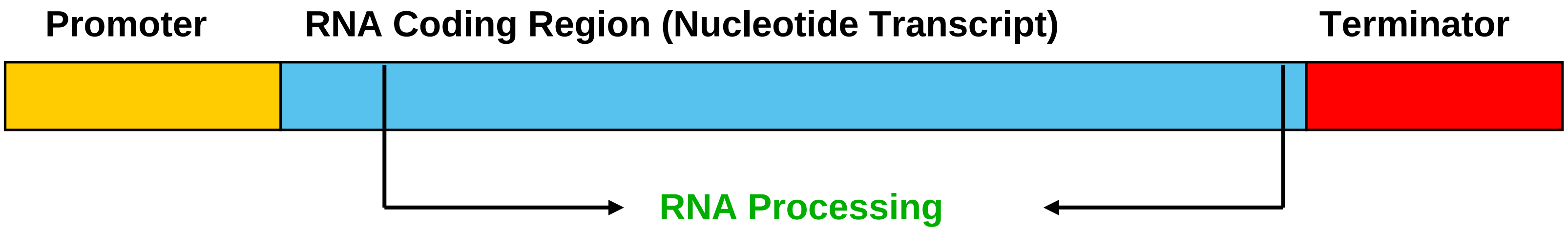
- John Chadwick,
  The Decipherment of Linear B,
  Cambridge Univ. Press, 1958

Microbial Protein and RNA Genes

Protein Gene

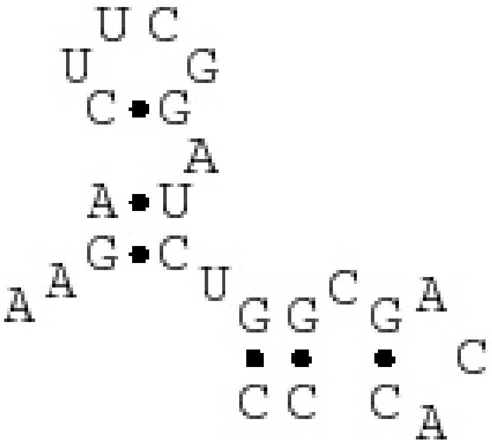

Promoter
Protein Coding Region (Triplets)
Terminator
Ribosome Binding Sequence (Shine-Dalgarno)
Start (ATG)
Stop (TGA)

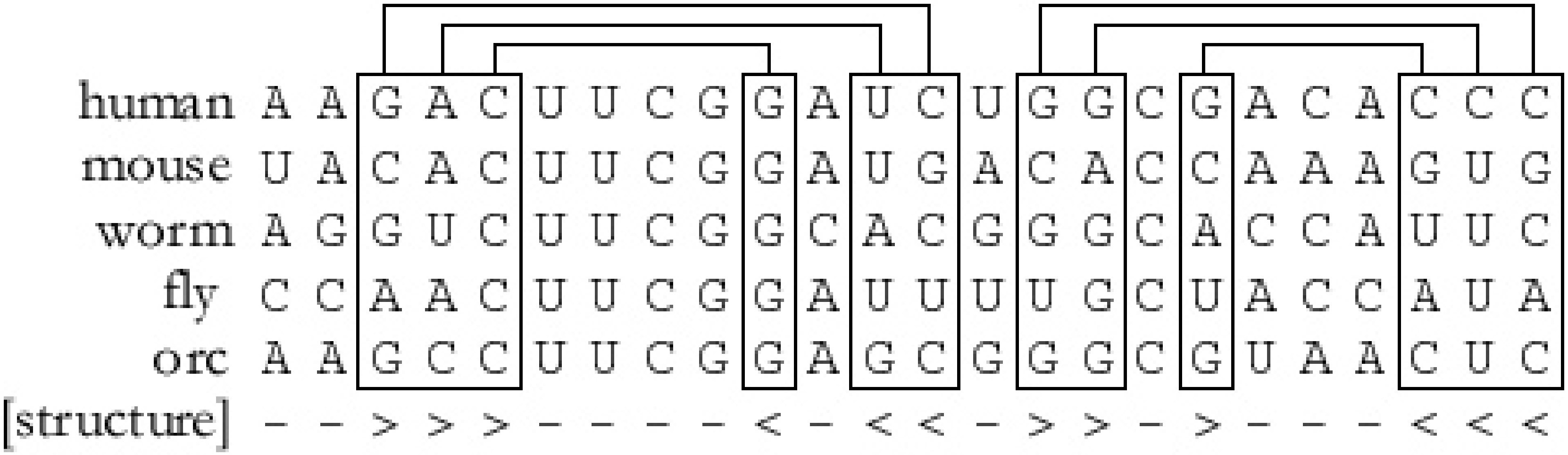
RNA Gene



- No ribosome binding sites
- No start or stop codons
- No triplet code

RNA structure: nested pairwise correlations





human	A	A	G	A	C	U	U	C	G	G	A	U	C	U	G	G	C	G	A	C	A	C	C	C
mouse	U	A	C	A	C	U	U	C	G	G	A	U	G	A	C	A	C	C	A	A	A	G	U	G
worm	A	G	G	U	C	U	U	C	G	G	C	A	C	G	G	G	C	A	C	C	A	U	U	C
fly	C	C	A	A	C	U	U	C	G	G	A	U	U	U	U	G	C	U	A	C	C	A	U	A
orc	A	A	G	C	C	U	U	C	G	G	A	G	C	G	G	G	C	G	U	A	A	C	U	C
[structure]	-	-	>	>	>	-	-	-	-	<	-	<	<	-	>	>	-	>	-	-	-	<	<	<

Context-free grammars

Noam Chomsky, 1956

Basic CFG
“production rules”

S → aS

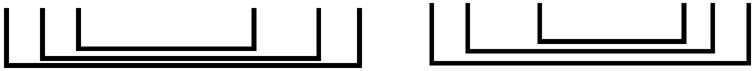
S → Sa

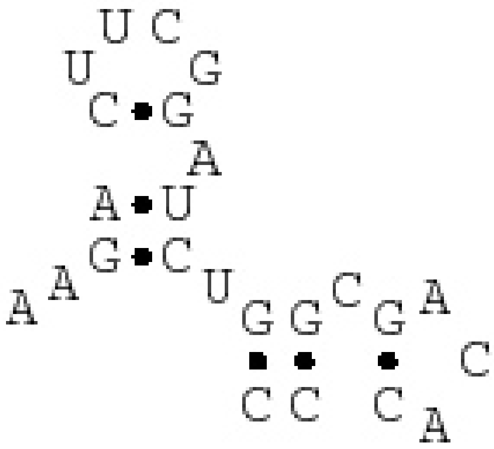
S → aSu

S → SS

a CFG “derivation”

S → aS
→ aaS
→ aaSS
→ aagScuS
→ aagaSucugSc
→ aagaSaucggScc
→ aagacSgaucuggcScc
→ aagacuSgaucuggcgSccc
→ aagacuuSgaucuggcgaSccc
→ aagacuucSgaucuggcgacSccc
→ aagacuucgSgaucuggcgacaSccc
→ aagacuucggaucuggcgacaccc



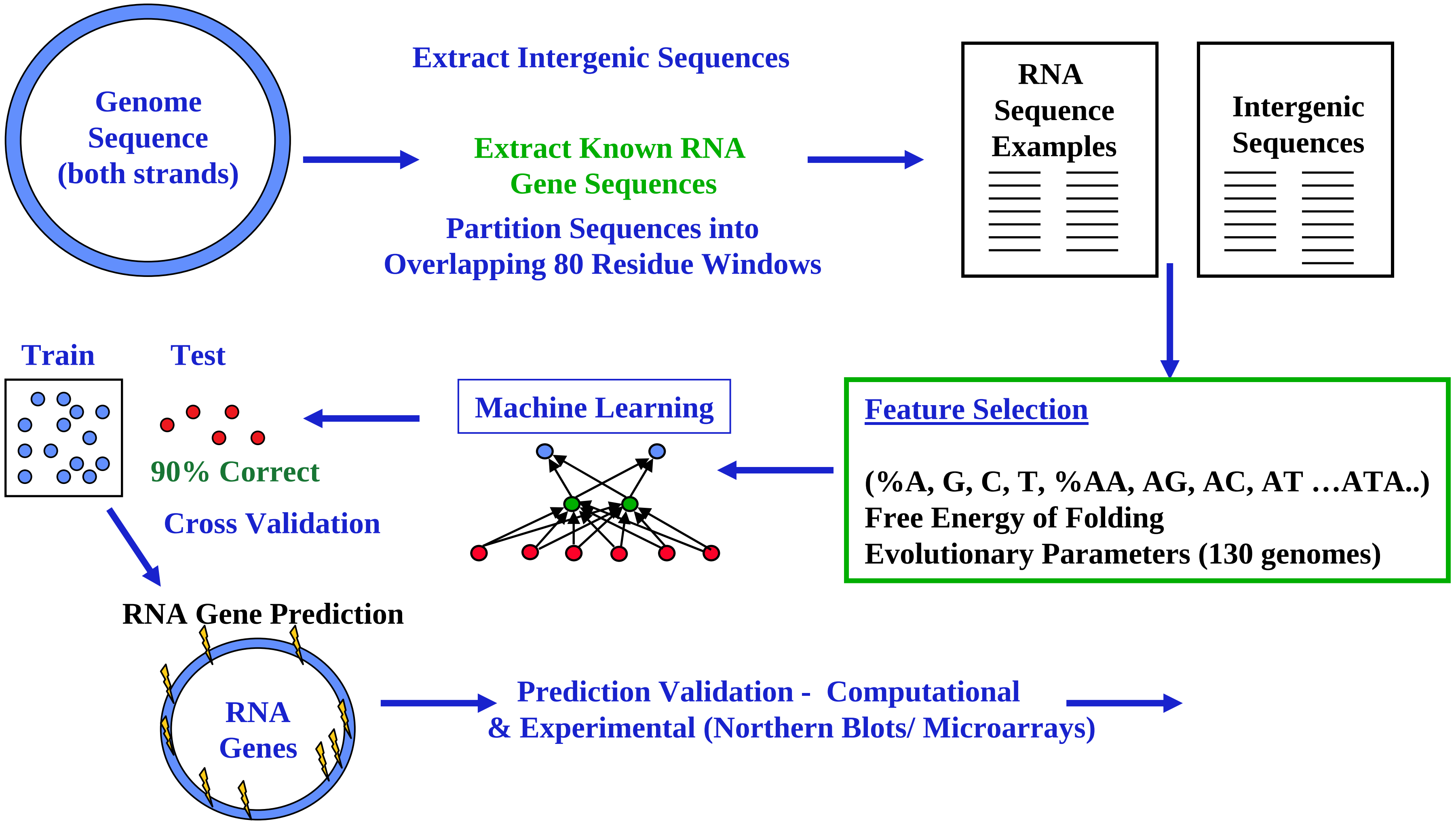


Machine Learning

Definition:

A computer program is said to learn **from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience.**

Flow Chart for RNA Gene Prediction



Genome Sequence (both strands)

Extract Intergenic Sequences

Extract Known RNA Gene Sequences

Partition Sequences into Overlapping 80 Residue Windows

RNA Sequence Examples

Intergenic Sequences

Feature Selection

(%A, G, C, T, %AA, AG, AC, AT …ATA..)
Free Energy of Folding
Evolutionary Parameters (130 genomes)

Machine Learning

Train

Test

90% Correct

Cross Validation

RNA Gene Prediction

RNA Genes

Prediction Validation - Computational & Experimental (Northern Blots/ Microarrays)

Learning from Partially Labeled Data

Determine initial negative set (N) such that :

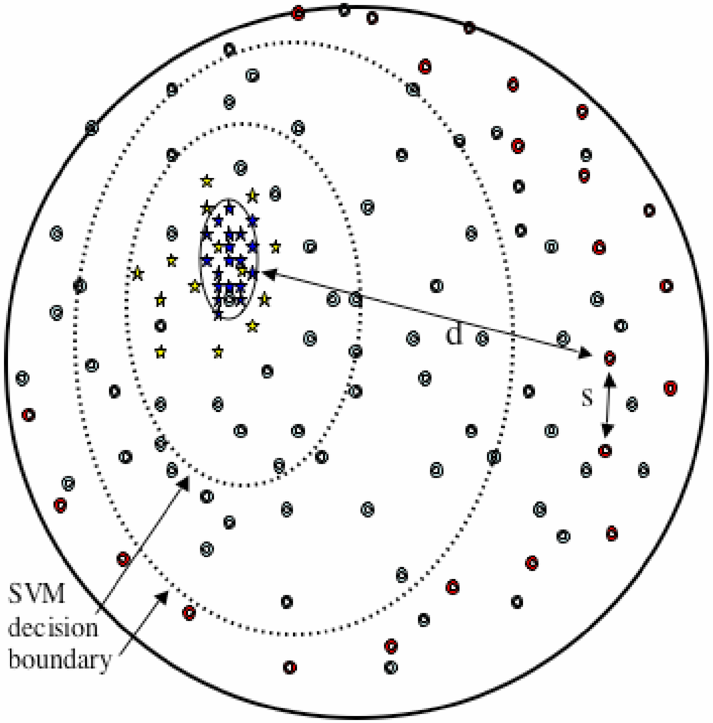
(1) Maximally distant from positive set (P)

(2) Maximally dissimilar from each other

$$(1)\ \max_{N\subset U} d(N,P), \qquad d(N,P)=\sum_{i\in N} d(x_i,P)$$

$$(2)\ \max_{N\subset U} d(N,N), \qquad d(N,N)=\sum_{i,j\in N} d(x_i,x_j)$$

$$\text{where} \qquad d(x_i,P)=\min_{j\in P}\left\|x_i-x_j\right\|$$

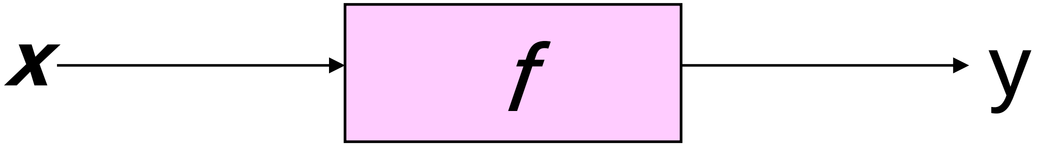


Solution is messy and expensive for lots of data.

Close enough and easy to compute

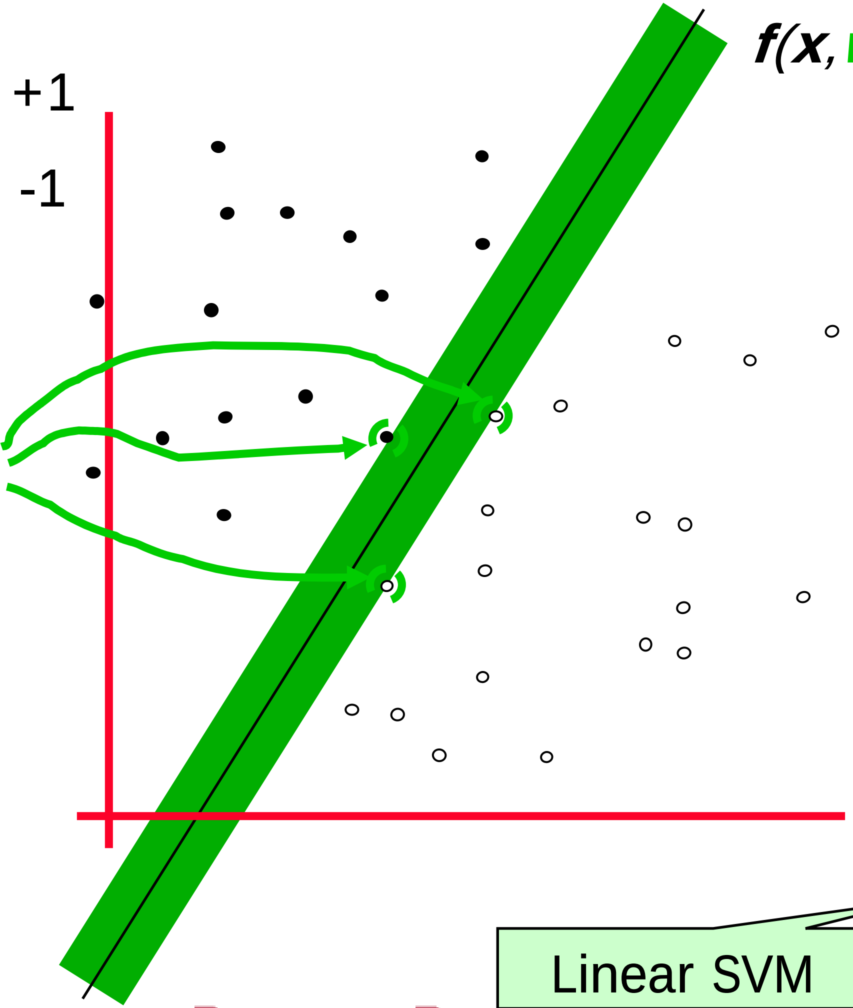
$$\max_{N\subset U}\left[d(N,P)d(N,N)\right] \longrightarrow \max_{i\subset(U-S)}\left[d(x_i,P)\sum_{j\in S} d(x_i,x_j)\right]$$

Maximum Margin Hyperplane



$f(\mathbf{x}, \mathbf{w}, b) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$

- ● denotes +1
- ○ denotes -1



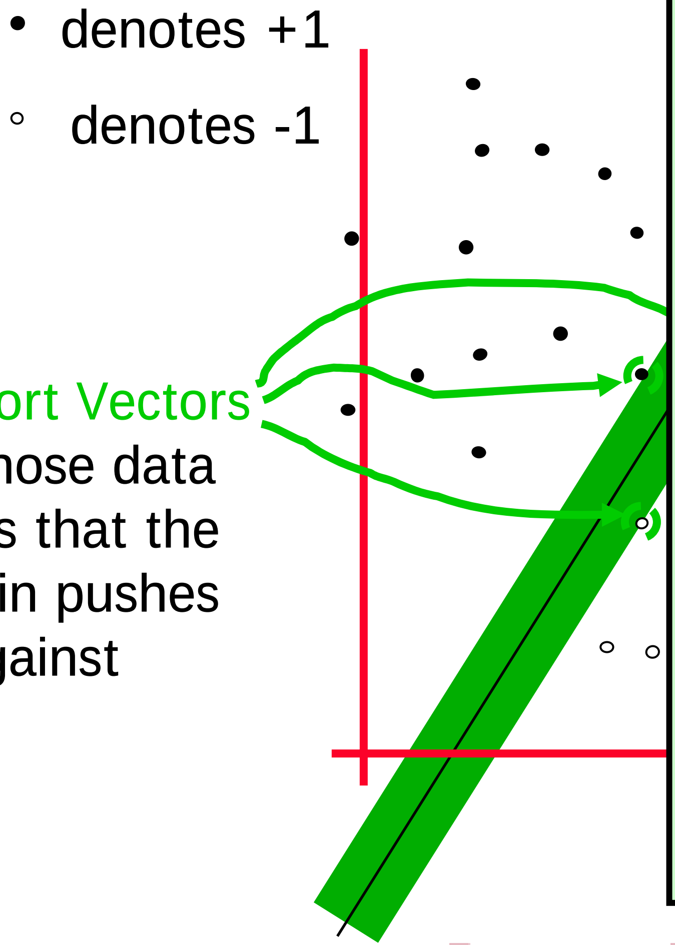
Support Vectors are those data points that the margin pushes up against

The maximum margin linear classifier is the linear classifier with the, um, maximum margin.

This is the simplest kind of SVM (Called an LSVM)

Linear SVM

Why Maximum Margin?



Support Vectors are those data points that the margin pushes up against

1. Intuitively this feels safest
2. If we've made a small error in the location of the boundary (it's been jolted in its perpendicular direction) this gives us least chance of causing a misclassification
3. Robust to outliers since the model is immune to change/removal of any non-support-vector data points
4. There's some theory (using VC dimension) that is related to (but not the same as) the proposition that this is a good thing
5. Empirically it works very well

SVM Kernel Functions

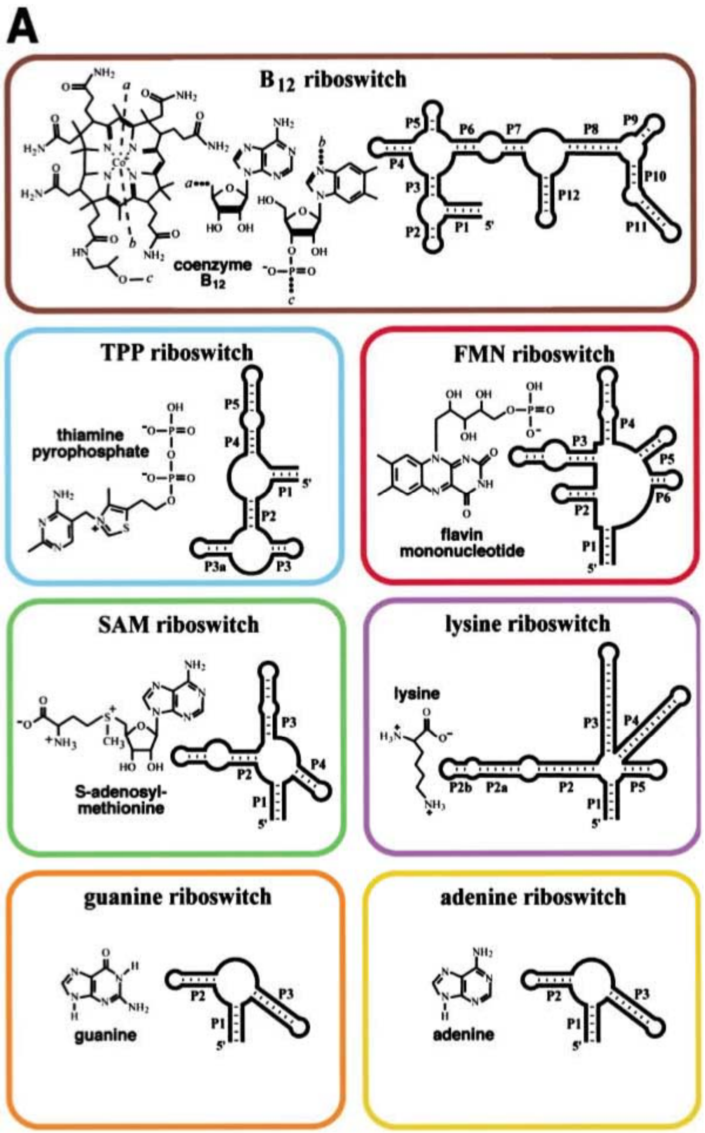
- ***K***(*a*,*b*)=(*a* **.** *b* +*1*)^d **is an example of an SVM kernel function**
- **Beyond polynomials there are other very high dimensional basis functions that can be made practical by finding the right kernel function**
  - **Radial-Basis-style Kernel Function:**

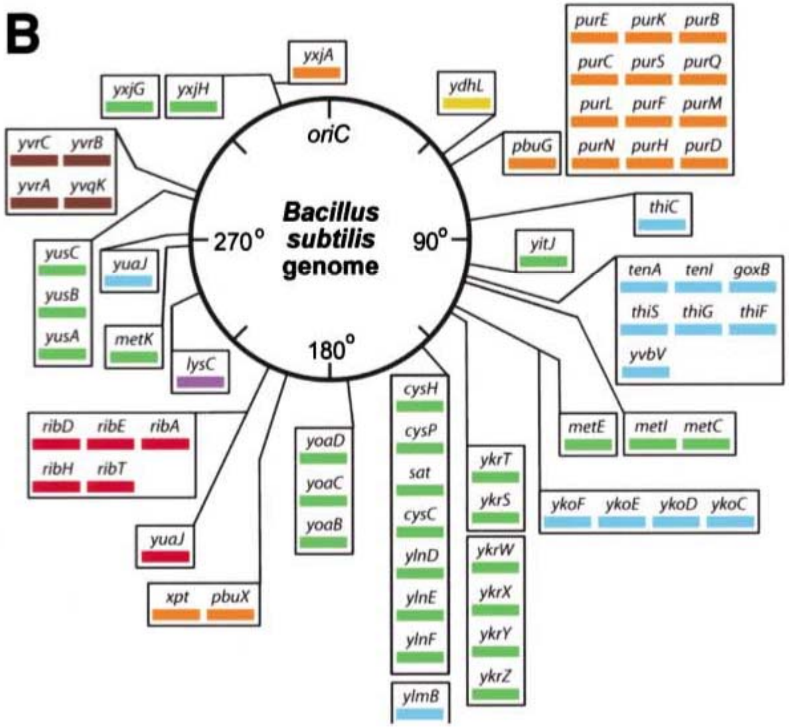
$$K(\mathbf{a},\mathbf{b}) = \exp\left(-\frac{\|\mathbf{a}-\mathbf{b}\|^2}{2\sigma^2}\right)$$

  - **Sigmoid Style Kernel Function:**

$$K(\mathbf{a},\mathbf{b}) = \tanh(\kappa\mathbf{a}.\mathbf{b}-\delta)$$

σ, κ and δ are magic parameters that must be chosen by a model selection method such as CV or VCSRM





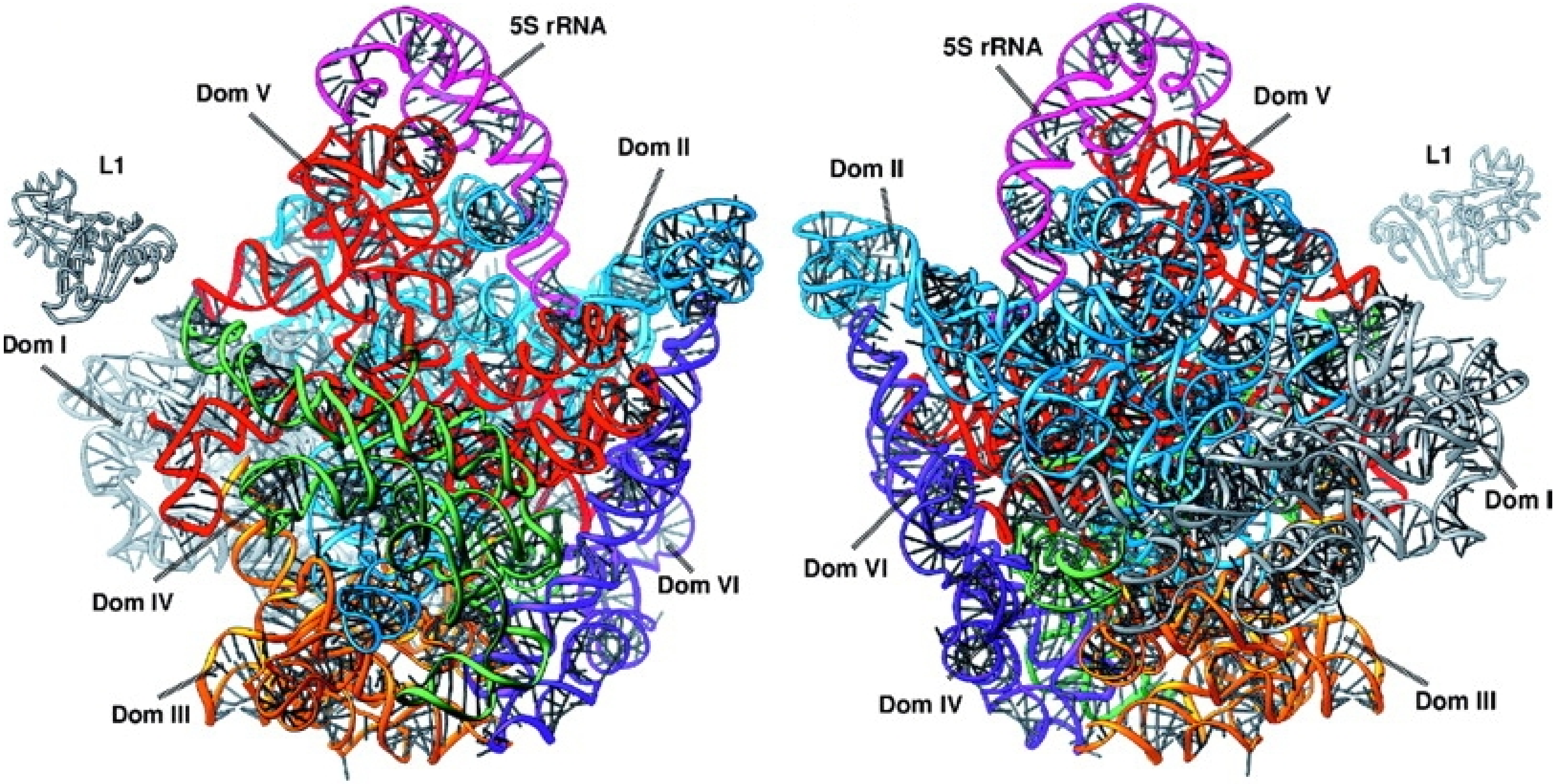
RIBOSWITCHES IN FUNDAMENTAL GENE CONTROL

A. THE SEVEN KNOWN RIBOSWITCHES AND THE METABOLITES THEY SENSE; NOTE THAT THE METABOLITES ALMOST ALL CONTAIN PYRIMIDINE OR PURINE MOIETIES.
B. GENETIC MAP OF *Bacillus subtilis* RIBOSWITCH REGULONS AND THEIR POSITIONS ON THE BACTERIAL CHROMOSOME; GENES ARE CONTROLLED BY RIBOSWITCHES OF MATCHING COLOR.

MANDAL *et al.* (2003), *Cell* **113**, 577-586

Structure of the large ribosomal subunit

Haloarcula marismortui
Ban et. al., Science 289:905, 2000



5S rRNA
Dom V
L1
Dom II
Dom I
Dom IV
Dom VI
Dom III
5S rRNA
Dom V
Dom II
L1
Dom I
Dom VI
Dom IV
Dom III